PERFORMANCE EVALUATION OF PERCEPTUAL QUALITY AND INTELLIGBILITY OF ENHANCED SPEECH USING MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

SYNOPSIS

Submitted for the award of the degree of

DOCTOR OF PHILOSOPHY

IN

BIOMEDICAL INSTRUMENTATION ENGINEERING

by

HEPSIBA D (Reg. No. 17PHEBP001)

Ph.D. Research scholar (Part Time)

Under the guidance of

Dr. JUDITH JUSTIN

Professor and Head

Department of Biomedical Instrumentation Engineering

School of Engineering

Avinashilingam Institute for Home Science and Higher Education for Women

Coimbatore

May 2022

PERFORMANCE EVALUATION OF PERCEPTUAL QUALITY AND INTELLIGIBILITY OF ENHANCED SPEECH USING MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

1. INTRODUCTION

In the process of communicating with each other, there are many methodologies such as sign language, Facial Expressions gestures, postures and many more. When the communication happens verbally, the information is shared by words through writing or speaking. Therefore, the most convenient and faster way of expressing one's feelings and thoughts is through speech. The communication does not end when the person speaks, it ends only when the speech is listened by the listener without any disturbance. Here comes the role of noise removal and enhancement in the speech signal which is prone to background noise and reverberation.

Speech enhancement [1, 2] is very important for any speech signal suffering from distortions, reflections and background noise that varies from place to place. Therefore, the speech enhancement techniques are crucial and very important for improving the speech quality in applications such as speaker recognition, automatic speech recognition (ASR) [3] and hearing aids [11]. Speech enhancement algorithms help in reducing noise without disturbing the quality of target speech.

The conventional algorithms of speech enhancement include minimum mean square error, spectral subtraction , Kalman filtering and iterative Wiener filtering. In the recent past, computational intelligence and machine learning have found wide applications in enhancing distorted speech and noise removal [4,5]. The latest trend in speech enhancement uses deep learning [13], which adopts the architecture of a deep neural network (DNN). As the process of speech enhancement is very important and it is the need of the hour in many applications the research work contributes to the enhancement of speech signal subjected to different types of noises by applying different algorithms and analyze its performance metrics in terms of quality and intelligibility.

In this research work, various algorithms for speech enhancement are implemented for the normal speech dataset and the alaryngeal speech dataset and the performance metrics of quality and intelligibility of the enhanced speech are analyzed.

2. LITERATURE REVIEW

In order to eliminate the non-stationary noises in the time domain, the noise components

are extracted from the corrupted speech signals and attenuated. The different DNN Based Speech Enhancement methodologies adopted vary based on neural network architecture, training the target and selection of training features. Nowadays, the deep learning models that are becoming popular in the field of speech enhancement are CNN[14], LSTM [15] which incorporate the transformation function to convert the spectral features of the noisy speech signal and clean speech signal.

The Back Propagation Network (BPN) in conjunction with DNN uses the perceptual weighted matrix to adjust the weight of the error matrix. In this system the error from the BPN was transferred to the first layer of DNN. The system tested for both stationary and non-stationary noise signals show significant difference in the results compared to the traditional DNN techniques [7]. Due to the regression function powerful modeling capacity is observed in DNN. The mapping function between the noisy and the clean speech signals is estimated using DNN and it depicts improvements in MMSE over the conventional methods [8]. The coded speech enhanced in the time domain using CNN is evaluated in narrowband and wideband speech codec. The input speech signal is processed in time and cepstral domain. The output signal is concatenated in time domain, and waveform reconstruction is performed in cepstral domain [12].

Long Short-Term Memory (LSTM) is a long-term temporal context system that has the ability to distinguish noisy and clean speech signals. Progressive learning used along with LSTM learns from intermediate target, specific to a SNR gain value. When the number of intermediate targets increase, the performance of LSTM-PL gets degraded. Hence, a densely connected LSTM is used by splicing the targets together for learning. The results inferred that LSTM yielded significant improvement in speech enhancement in terms of gain for all the noise types and better performance is attained for more targets [9].

3. RESEARCH OBJECTIVE

The main goal of the research is to enhance the quality and intelligibility of speech for speech disorder patients. People with speech disorders experience difficulties like distorting sounds, adding sounds or syllables to words, rearranging syllables, having difficulty pronouncing words correctly, struggling to say the correct word or sound, speaking with a hoarse or raspy voice, speaking very softly. Therefore, enhancing the speech results in efficient speech recognition.

4. SPEECH ENHANCEMENT SYSTEM

The objective of speech enhancement is to improve the intelligibility and perceptual quality of degraded speech using machine learning and deep learning algorithms.

4.1 MODEL OF SPEECH SIGNAL

The noisy speech signal is acquired from adding the clean speech signal with the different types of noise as given in equation 1. The objective of the research work is to retrieve the clean speech signal from the noisy speech signal by eliminating the noise.

$c(n)$: Clean Speech Signal

$b(n)$: Noise Signal

$s(n)$: Noisy Speech Signal

$$s(n) = c(n) + b(n) \quad (1)$$

4.2 DATASET

The clean speech signal is taken from the University of Edinburgh, Centre for Speech Technology Research (CSTR) (https://datashare.is.ed.ac.uk/handle/10283/2791). The dataset contains nearly 400 speech sentences. These speech sentences are taken for training and different types of noise are added with different decibels. The dataset is divided into training and testing data by applying holdout validation method. 80% of the dataset is taken as training data and 20% is taken as the testing data. The noisy speech signal dataset is created by mixing different noises such as street noise, car noise, washing machine noise, train-whistle, jet plane and rainbow noise with the clean speech signal at different noise levels 0 dB, 5 dB, 10 dB and 15 dB. The noisy speech signal taken for testing the network includes the noises such as Babble Noise, Restaurant Noise, Airport Noise, Exhibition Noise to check the network's denoising ability.

4.3 FEATURE EXTRACTION

The first step is to convert the speech signal from the time domain to the frequency domain using STFT to extract features. The magnitude STFT vectors of the clean speech and the noisy speech are input features to the machine learning and deep learning algorithms. Therefore, the speech signal is divided into a 10 ms frame with no frameshift. In converting from the time domain to frequency domain using STFT, the hamming window is utilized with a window length of 256 samples and 75% overlap. For training and testing purposes, the speech signal is down sampled to

an 8 kHz signal and a 256-point FFT is implemented and the number of frequency bins is 129. The clean speech corpus taken from the open-source dataset was contaminated by the noise signals at different noise levels.

4.4 PERFORMANCE METRICS

- **Signal to Noise Ratio (SNR)**

$$SNR_{dB} = 20 \log_{10} \frac{S_rms}{N_rms} \quad (2)$$

Where S_rms – root mean square of speech signal

N_rms - root mean square of level of noise

- **Perceptual Evaluation of Speech Quality (PESQ)**

PESQ is a subjective quality measurement and it is based on the mean opinion score based on the evaluation given by the listeners and standardized by International Telecommunications Union (ITU). The PESQ values are based on the quality on a scale of 0 (Worst) to 5 (Best). The higher the value of PESQ indicates good quality of the speech signal.

- **Short Time Objective Intelligibility (STOI)**

STOI is a subjective intelligibility measurement, larger the value better the speech intelligibility. The STOI value ranges between 0-1.

5. SPEECH ENHANCEMENT ALGORITHMS

5.1 WIENER FILTER

The presence of noise is unavoidable in real-world scenarios of speech processing. The most fundamental methodology in noise reduction of a speech signal is the optimal Wiener filter. The Wiener filter acts as a linear filter that could be utilized to separate the clean speech signal from the noisy speech signal by reducing the MSE between the estimated signal and the original signal. As the Wiener filter can achieve noise reduction, it also has the disadvantage of losing the speech signal's integrity. Therefore, the speech misrepresentation should be managed in such a way by adequately manipulating the Wiener filter or to have explicit knowledge of the speech signal. In any speech communication system, the speech signal could be distorted by background

noise and reverberations. Therefore, noise reduction methodologies and speech enhancing techniques are needed to obtain the desired speech signal from the corrupted ones.

5.2 BACK PROPAGATION NETWORK

Back Propagation Network (BPN) is a feed-forward neural network with transfer functions like tan-sigmoid. These transfer functions are used in both the hidden and the output layer [13]. The mathematical expression for the calculating the output of the BPN network is as follows. Considering the input vectors, $I = [p_1, p_2, \dots\dots\dots, p_n]$, the output of the hidden layer can be calculated as,

$$n = IW + b \tag{3}$$

$$a = f(n) = \frac{2}{\left(1+e^{(-2*n)}\right)^{-1}} \tag{4}$$

W is the weight vector, and b is the bias input. 300 speech utterances are taken to train the BPN. Eqn. 3 and 4 shows the mathematical expression for the calculation of output in BPN network architecture. Errors are calculated as the difference between the target output and the actual output of the network. To minimize the sum of these errors, the target is to minimize the average difference between them.

$$mse = \frac{1}{M}\sum_{k=1}^{M} e(k)^2 = \frac{1}{M}\sum(t(k) - a(k))^2 \tag{5}$$

$a(k)$ is the actual output, $t(k)$ represents the target output, and mse represents Mean Square Error. By updating weights and bias values based on the goal average error value, the algorithm gets more accurate.

5.3 DEEP CONVOLUTIONAL NEURAL NETWORK (DEEP CNN)

CNN is designed with multiple hidden layers with rectified linear unit (ReLU) activation function for speech enhancement. The input applied to the Deep CNN system is the frames of the noisy speech signal, and the expected output is the denoised speech signal. The clean and noisy speech signal is converted to the frequency domain using STFT. The regression network uses the magnitude of the noisy speech signal to reduce the mean square error between the denoised speech signal and the clean speech signal.

The output from the Deep CNN gives the denoised signal in the frequency domain. The denoised speech signal is converted to the time domain using the output magnitude spectrum from

the Deep CNN network and the phase of the noisy speech signal. The denoising algorithm utilizes convolutional layers in which each neuron is connected to all the activations in the previous layer. Deep CNN is used to learn the spectral mapping from the noisy speech signal to the clean speech signal.

5.4 LONG SHORT TERM MEMORY (LSTM) NETWORK

LSTM comprises three distinct components: the forget gate, input gate, and output gate. LSTMs acknowledge and remember information entering a network and discard the unnecessary information. Layer inputs are taken from the previous layer and the presentation layer. The presentation layer determines the input layer state. Un-important information is removed from the network state by utilizing the forget state feature. Sequential data, like time and audio data, are analyzed using the LSTM network.

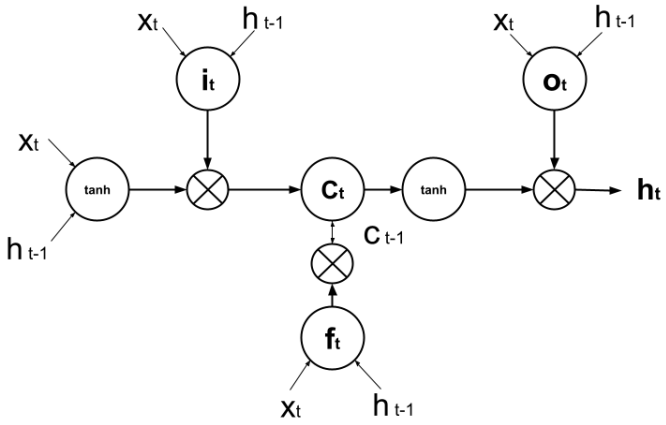

Fig 1 LSTM Architecture

A single block LSTM cell is shown in Fig. 1. There are several key components, such as the memory cell state C_t, the input gate i_t, the forget gate f_t, the output gate o_t and the peepholes. x_t and h_t represent the input and output vectors of the block at time t, respectively. Multiplication of two vectors ⊗ multiplies them by a point-wise formula. As for the non-linear activation functions, the logistic sigmoid function is employed in every gate. As a result, the network can dynamically decide what information to update, store, discard, and output. Therefore, it will efficiently utilize temporal information. For the subsequent separation operation, the LSTM may well capture the inherent statistical properties of speech and noise, particularly under non-stationary noise conditions [10].

5.5 RESULTS AND DISCUSSION

Table 1 Evaluation of Performance Metrics

		SNR						PSEQ					STOI				
		Noise Level (dB)	Noisy Signal	Wiener	BPN	Deep CNN	LSTM	Noisy Signal	Wiener	BPN	Deep CNN	LSTM	Noisy Signal	Wiener	BPN	Deep CNN	LSTM
SEEN NOISE	Washing Machine Noise	0	19.92	34.78	26.97	30.58	31.90	1.56	1.73	1.85	2.20	2.50	0.53	0.02	0.57	0.60	0.60
		5	24.92	35.35	28.00	33.43	36.15	1.74	2.07	2.06	2.49	2.68	0.58	0.03	0.60	0.68	0.72
		10	29.48	36.54	29.63	34.85	37.30	1.96	2.17	2.24	2.70	3.15	0.59	0.02	0.63	0.75	0.71
		15	30.92	38.33	32.83	35.74	38.39	2.07	2.39	2.55	2.92	3.56	0.63	0.05	0.67	0.81	0.79
	Rainbow Noise	0	19.98	35.29	26.99	30.57	32.89	1.56	1.65	1.80	1.99	2.21	0.51	0.00	0.51	0.60	0.58
		5	24.84	36.71	28.32	32.56	35.28	1.83	1.97	2.10	2.39	2.51	0.58	0.05	0.61	0.71	0.76
		10	29.59	37.94	31.51	34.66	37.11	2.02	2.20	2.29	2.56	2.85	0.61	0.03	0.62	0.74	0.77
		15	30.83	38.44	33.84	35.52	38.17	2.19	2.21	2.56	2.88	3.70	0.61	0.07	0.64	0.80	0.86
	Babble Noise	0	19.68	33.46	26.91	29.26	32.48	1.54	1.65	1.78	2.14	2.49	0.47	0.02	0.51	0.55	0.69
		5	24.68	34.37	27.95	32.45	34.37	1.72	1.82	1.90	2.34	2.88	0.54	0.03	0.58	0.61	0.65
		10	27.24	34.69	29.57	33.00	35.81	1.84	2.02	2.28	2.46	3.25	0.55	0.02	0.61	0.63	0.73
		15	30.68	36.38	32.78	33.79	36.30	1.99	2.16	2.40	2.97	3.56	0.59	0.05	0.62	0.66	0.78
	Airport Noise	0	19.70	35.97	28.08	30.25	31.20	1.67	1.80	1.99	2.22	2.74	0.60	0.02	0.62	0.51	0.73
		5	24.62	36.03	30.72	31.58	32.24	1.78	1.92	2.20	2.54	2.94	0.64	0.05	0.67	0.75	0.78
		10	29.32	37.09	32.55	33.81	34.53	1.85	2.05	2.49	2.88	3.47	0.65	0.02	0.69	0.79	0.86
		15	30.91	38.49	33.83	34.57	35.25	1.90	2.22	2.57	3.23	3.96	0.66	0.07	0.70	0.83	0.91
UNSEEN NOISE	Jet Plane Noise	0	19.99	36.45	27.01	33.29	37.61	1.27	1.56	1.65	2.17	2.49	0.49	0.07	0.55	0.67	0.71
		5	24.82	38.99	29.99	34.24	38.96	1.71	1.97	2.17	2.39	2.82	0.56	0.07	0.61	0.71	0.77
		10	29.56	42.10	30.05	35.13	39.58	1.81	2.08	2.32	2.68	3.36	0.57	0.02	0.61	0.77	0.86
		15	33.73	44.81	31.01	36.11	40.76	2.09	2.27	2.54	2.72	3.94	0.58	0.04	0.62	0.77	0.89
	Street Noise	0	19.74	34.65	26.93	27.93	32.33	1.44	1.66	1.94	2.12	2.50	0.47	0.06	0.49	0.56	0.61
		5	24.60	35.75	28.26	30.60	34.81	1.81	1.97	2.15	2.49	2.80	0.54	0.06	0.59	0.64	0.74
		10	29.35	36.24	30.45	31.96	35.64	1.90	2.30	2.44	2.79	3.34	0.57	0.03	0.60	0.70	0.77
		15	30.59	37.54	31.78	33.62	38.23	2.17	2.32	2.60	2.98	4.19	0.57	0.05	0.62	0.70	0.78
	Train Whistle Noise	0	19.94	37.80	29.13	34.77	35.23	1.79	1.89	2.04	2.30	2.44	0.64	0.01	0.64	0.70	0.79
		5	24.86	38.10	30.77	36.86	38.27	1.90	2.12	2.35	2.53	2.75	0.68	0.05	0.69	0.82	0.76
		10	29.56	38.98	32.61	37.01	40.56	2.07	2.32	2.54	2.79	3.08	0.69	0.08	0.71	0.78	0.83
		15	31.15	39.35	33.89	37.92	42.28	2.22	2.36	2.63	2.98	3.77	0.70	0.02	0.72	0.89	0.90
	Restaurant Noise	0	19.75	33.33	26.96	30.61	32.97	1.25	1.43	1.70	2.24	2.59	0.45	0.02	0.53	0.60	0.63
		5	24.58	34.77	29.94	32.62	35.46	1.79	1.88	1.96	2.63	2.91	0.52	0.07	0.59	0.65	0.70
		10	29.32	35.39	30.00	34.11	37.34	2.09	2.26	2.57	3.05	3.45	0.53	0.09	0.59	0.72	0.78
		15	33.49	36.59	30.95	35.67	38.38	2.27	2.41	2.79	3.34	4.03	0.53	0.03	0.60	0.76	0.81

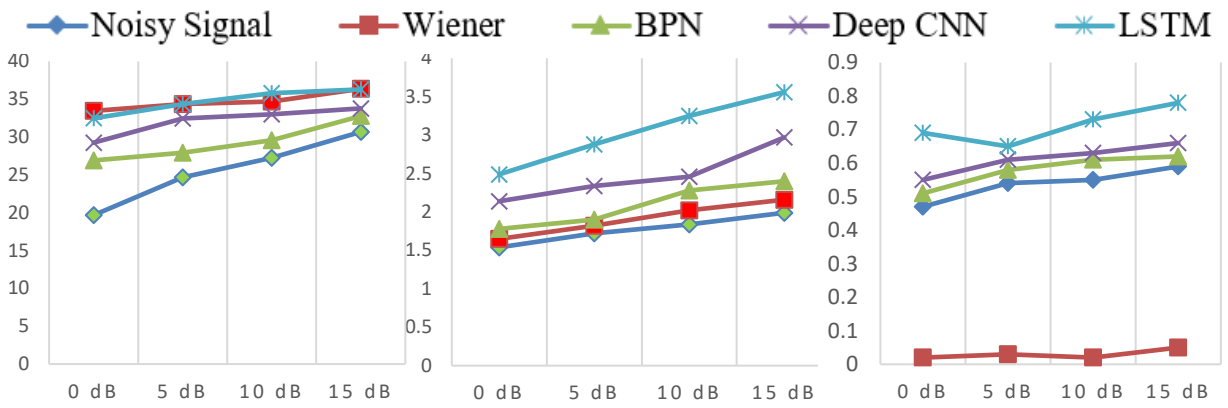

Fig 2 Performance analysis for Babble Noise

Speech Quality is determined based on SNR and PESQ, and Intelligibility is determined based on STOI these performance metrics are shown in Table 1. The Speech Enhancement Systems are trained with different noises such as Washing Machine Noise, Rainbow Noise, Babble Noise and Airport Noise for different values of noise levels such as 0dB, 5dB, 10dB and 15dB. For testing purposes, the model is subjected to the unseen noises such as Jet Airplane Noise, Street Noise, Train Whistle Noise and Restaurant Noise at different noise levels such as 0dB, 5dB, 10dB and 15dB. The performance of the Speech Enhancement methods were evaluated and analyzed. The SNR of Wiener Filter outperforms all the other methods. But the PESQ of the LSTM network is good compared to the other algorithms. Wiener Filter does not have the capability to improve the Intelligibility therefore it is observed from the results that STOI of machine learning and deep learning based Speech Enhancement performs better than Wiener Filtering Method. The performance of BPN is better compared to the Wiener filter but less compared to the Deep CNN and LSTM as shown in Fig 2.

6. SPEECH DISORDER CAUSED BY LARYNGECTOMY

Sound is created by swallowing air and gargling, which converts air into words. Larynx is the portion of the throat that houses the vocal cords, that allows a person to produce sound. When the larynx or the voice box becomes infected with cancer, the vocal folds or the larynx will be removed surgically and it is referred to as total laryngectomy. Patients who have undergone total laryngectomy will only be able to speak in whispers due to the removal of the larynx responsible for producing vocal vibrations. With prostheses or specialized speech therapy methods, patients who undergo partial or complete laryngectomy are usually incapable of speaking more than hoarse

whispers. A small hole or puncture is made in the common wall between the trachea and esophagus. This puncture is visible inside the stoma and becomes a path to allow airflow into the esophagus. A small silicone one-way valve is placed inside the puncture to keep it open. This one-way valve is called a tracheoesophageal prosthesis.

6.1 SPEECH ENHANCEMENT SYSTEM FOR ALARYNGEAL SPEECH

Speech enhancement system for alaryngeal speech as shown in Fig 3, helps in improving the quality of speech for people who have undergone laryngectomy as their voice is very feeble and hoarse. Due to this reason, it is challenging for others to understand the sentences spoken by the total laryngectomy patients with Blom Singer non-indwelling voice prosthesis. The speech spoken by the laryngectomy patients are referred as alaryngeal speech. The various speech enhancement methodologies are adopted for enhancing the alaryngeal speech that is subjected to various types of noises in real-time.

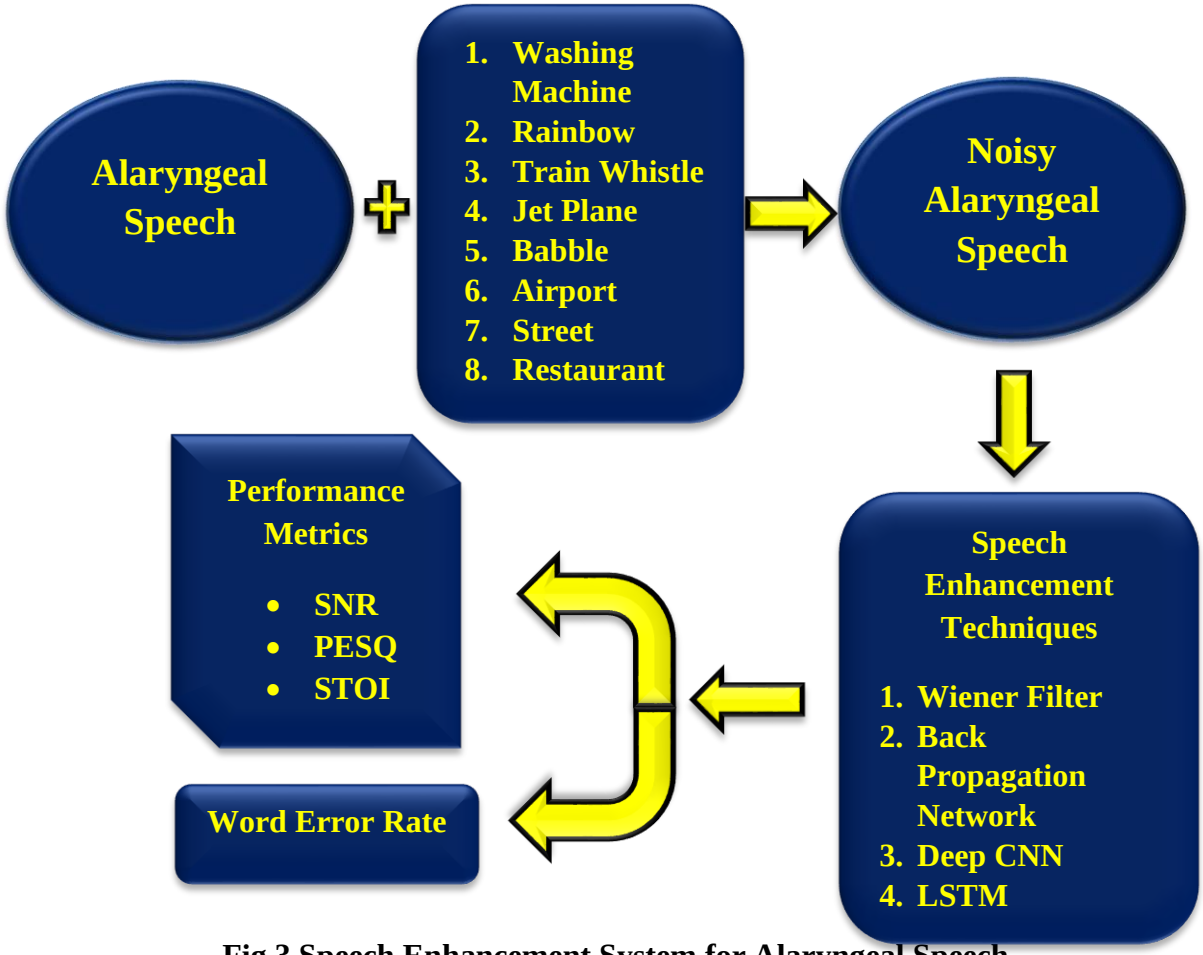

Fig 3 Speech Enhancement System for Alaryngeal Speech

6.1 DATASET FOR ALARYNGEAL SPEECH

The alaryngeal speech dataset is created by recording the IEEE sentences spoken by the patient who have undergone total laryngectomy. The clean alaryngeal speech is mixed with different noise types such as Washing Machine Noise, Rainbow Noise, Trainwhistle Noise, Jet Airplane Noise, Babble Noise, Street Noise, Airport Noise and Restaurant Noise at different noise levels such as 0dB, 5dB,10dB and 15dB.

6.2 SPEECH ENHANCEMENT SYSTEM FOR ALARYNGEAL SPEECH

Speech enhancement helps in improving the quality of speech for people who have undergone laryngectomy as their voice is very feeble and hoarse. Due to this reason, it is challenging for others to understand the sentences spoken by the total laryngectomy patients with Blom Singer non-indwelling voice prosthesis. The speech spoken by the laryngectomy patients are referred as alaryngeal speech. The various speech enhancement methodologies are adopted for enhancing the alaryngeal speech that is subjected to various types of noises in real-time.

6.3 EXPERIMENTAL RESULTS & DISCUSSIONS

For different noise levels and different types of noise, the Wiener filtering technique shows substantial improvement in the SNR of the denoised alaryngeal speech over the Deep CNN model, and LSTM. Wiener Filter, the best conventional filter for speech enhancement, performs well in terms of SNR values compared to the Machine learning and Deep learning models as shown in Table 2. The intelligibility of speech is compromised in Wiener filter which reduces its clarity and it does not lead to an improvement in intelligibility based on the STOI scores.

BPN performs reasonably well over Wiener filters but not better than Deep CNN and LSTM networks. The PESQ values of BPN are moderate, while the SNR values are good. Deep CNNs shows drastic result when it comes to STOI values, which represent the intelligibility of a denoised speech signal. In terms of LSTM metrics, notable performance is seen in terms of SNR and PESQ for all noise types and at varying noise levels. LSTM also achieves excellent results compared to other methods based on STOI values, improving speech intelligibility.

6.4 WORD ERROR RATE

In order to validate the performance of the speech enhancement methods, the IEEE speech sentences spoken by laryngectomy patients are analyzed by calculating the Word Error Rate (WER). The WER is 100% when the alaryngeal speech is subjected to noise and it is improved

when applying Speech Enhancement Algorithms. The WER is shown in Fig 4 for the speech signal enhanced by various algorithms are analyzed. It is apparent that the alaryngeal speech signal enhanced by the LSTM algorithm results in less word error rate and increases the chances of recognition of speech sentences. The word error rate is 18% and the reason behind this is the difficulty faced by the laryngectomy patients in pronouncing the sentences with the voice prosthesis device.

Table 2 Performance Evaluation of Enhanced Alaryngeal Speech

			SNR					PESQ					STOI				
		Noise Level	Noisy Signal	Wiener Filter	BPN	Deep CNN	LSTM	Noisy Signal	Wiener Filter	BPN	Deep CNN	LSTM	Noisy Signal	Wiener Filter	BPN	Deep CNN	LSTM
Seen Noises	Washing Machine Noise	0 dB	18.09	33.39	29.18	33.06	34.59	1.35	1.68	1.81	2.11	2.48	0.45	0.02	0.45	0.45	0.58
		5 dB	23.09	35.39	34.09	35.47	36.49	1.46	1.58	2.02	2.02	2.75	0.54	0.03	0.55	0.54	0.68
		10 dB	27.65	36.92	35.72	37.92	38.61	1.57	1.99	2.06	2.38	2.94	0.62	0.02	0.62	0.65	0.75
		5 dB	32.00	37.51	36.02	38.12	41.70	1.62	1.83	2.31	2.90	3.62	0.67	0.05	0.73	0.81	0.86
	Rainbow Noise	0 dB	18.15	32.45	29.36	33.07	34.45	1.31	1.33	1.62	1.70	2.30	0.42	0.00	0.46	0.62	0.68
		5 dB	23.00	36.93	32.57	34.41	37.72	1.56	1.67	1.78	2.07	2.89	0.53	0.05	0.56	0.62	0.66
		10 dB	27.75	36.75	35.59	36.86	40.15	1.60	1.75	1.98	2.44	3.24	0.63	0.03	0.65	0.69	0.74
		15 dB	32.00	40.81	35.91	37.34	42.73	1.64	1.88	1.95	2.70	4.11	0.65	0.07	0.66	0.71	0.77
	Babble Noise	0 dB	19.08	33.06	30.78	34.59	35.21	1.37	1.70	1.83	2.26	2.72	0.45	0.02	0.46	0.55	0.62
		5 dB	22.65	33.97	32.84	34.71	36.37	1.48	1.60	2.04	2.51	2.95	0.58	0.03	0.60	0.63	0.68
		10 dB	26.68	35.63	35.29	37.26	38.49	1.60	2.01	2.19	2.88	3.55	0.69	0.02	0.71	0.74	0.75
		15 dB	27.19	37.13	34.94	37.41	39.90	1.64	1.85	2.33	3.20	4.22	0.70	0.05	0.76	0.81	0.81
	Airport Noise	0 dB	18.12	33.88	31.78	34.60	38.64	1.48	1.68	1.96	2.20	2.72	0.48	0.01	0.52	0.58	0.79
		5 dB	23.04	34.64	33.15	36.02	39.06	1.54	1.74	1.94	2.57	3.17	0.62	0.06	0.62	0.65	0.82
		10 dB	25.33	35.44	34.80	37.39	41.52	1.60	1.94	2.54	2.77	3.80	0.66	0.08	0.67	0.72	0.85
		15 dB	30.15	39.52	39.06	40.91	42.09	1.65	2.10	2.33	2.79	4.09	0.76	0.02	0.78	0.77	0.89
Unseen Noises	Jet Plane Noise	0 dB	18.15	35.08	31.92	34.10	37.09	1.26	1.46	1.61	2.15	2.43	0.41	0.07	0.53	0.64	0.67
		5 dB	22.99	35.08	34.03	36.58	38.78	1.52	1.60	1.77	2.36	2.99	0.58	0.07	0.64	0.72	0.76
		10 dB	27.73	36.10	37.35	39.41	46.33	1.65	1.78	1.86	2.89	3.69	0.57	0.02	0.70	0.80	0.85
		15 dB	31.89	39.54	36.14	40.97	47.66	1.88	1.93	2.28	3.12	3.99	0.64	0.04	0.72	0.82	0.86
	Street Noise	0 dB	17.75	31.73	28.71	32.69	34.07	1.33	1.36	1.64	1.91	2.64	0.52	0.01	0.58	0.64	0.66
		5 dB	22.62	34.63	32.03	37.09	37.15	1.58	1.69	1.81	2.37	3.42	0.64	0.05	0.72	0.77	0.82
		10 dB	25.76	35.62	34.21	36.77	38.26	1.63	1.78	2.00	2.64	3.85	0.75	0.03	0.77	0.80	0.85
		15 dB	32.14	38.97	37.55	40.98	42.91	1.66	1.90	2.29	2.98	4.23	0.78	0.07	0.80	0.86	0.89
		0 dB	18.10	35.32	34.58	35.22	38.61	1.46	1.65	1.93	2.19	2.41	0.61	0.01	0.62	0.68	0.72
		5 dB	23.03	36.01	35.86	37.35	38.04	1.51	1.71	1.91	2.49	2.87	0.67	0.05	0.71	0.75	0.82

	Train Whistle Noise	10 dB	27.72	38.10	36.23	38.19	41.07	1.58	1.91	2.51	2.77	3.40	0.74	0.08	0.80	0.81	0.85
		15 dB	32.06	37.08	35.69	41.52	44.75	1.63	2.07	2.31	2.99	3.87	0.79	0.02	0.83	0.88	0.89
	Restaurant Noise	0 dB	17.75	31.22	34.29	34.31	37.28	1.29	1.48	1.64	2.10	2.39	0.46	0.07	0.49	0.53	0.59
		5 dB	21.62	35.34	32.01	36.48	38.84	1.55	1.63	1.79	2.46	2.76	0.58	0.07	0.59	0.62	0.71
		10 dB	25.12	36.63	33.84	39.94	41.97	1.68	1.91	2.28	2.89	3.36	0.69	0.02	0.72	0.76	0.77
		15 dB	27.18	37.22	34.92	40.62	42.62	1.71	2.15	2.50	3.24	3.92	0.76	0.04	0.78	0.84	0.89

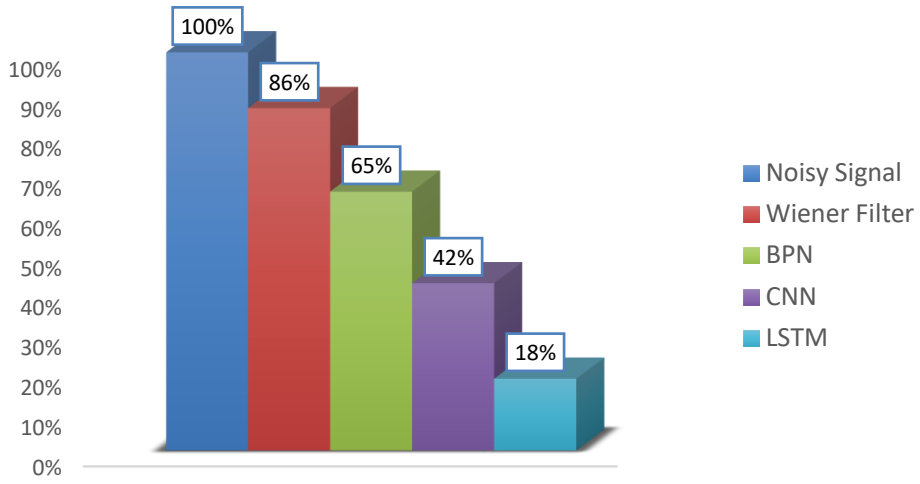

Fig 4 Validation of Speech Enhancement by Word error rate

7. CONCLUSION

As the number of laryngectomy patients increases every year, the voice prosthesis device provides an easy way for patients to recover and start conversing after surgery. This research work indicates that alaryngeal speech enhancement could significantly enhance communication effectiveness for laryngectomy patients by utilizing existing speech enhancement technologies. A comparison of four existing speech enhancement methods is undertaken in this research work. An effort is made towards denoising and enhancing alaryngeal speech signals using Wiener filter, BPN, Deep CNN and LSTM networks. Based on the results, it is evident that the Wiener filter performs well in terms of SNR, but the BPN gives intermediate results in terms of SNR, PESQ and STOI. However, Deep CNN shows improved performance in terms of PESQ and STOI. The LSTM shows excellent SNR, PESQ, STOI and WER results. Thus, it could be noted that the LSTM outperforms all other speech enhancement models**.**

8. ORGANIZATION OF THIS THESIS

Chapter 1 discusses about the motivation, objective and necessity of the research work. Chapter 2 interprets the literature review of the research work which focuses on the available concepts in the literatures and related works. Chapter 3 focuses on the traditional methodology of speech enhancement. Chapter 4 describes the design and implementation of the machine learning and deep learning algorithms for speech enhancement. Chapter 5 emphasis on the speech generation and speech disorders caused by Laryngectomy. Chapter 6 deals with the speech enhancement techniques for laryngectomy patients. Chapter 7 summarizes the entire research work and concludes the best algorithm for enhancement of alaryngeal speech.

REFERENCES

[1] Makino, S., & Chen, J. (2005). *Speech Enhancement*. Springer Science & Business Media.

[2] Loizou, P. C. (2007). *Speech enhancement: theory and practice*. CRC press.

[3] Kawase, T., Okamoto, M., Fukutomi, T., & Takahashi, Y. (2020). Speech enhancement parameter adjustment to maximize accuracy of automatic speech recognition. *IEEE Transactions on Consumer Electronics*, *66*(2), 125-133.

[4] Sekiguchi, K., Bando, Y., Nugraha, A. A., Yoshii, K., & Kawahara, T. (2019). Semi-supervised multichannel speech enhancement with a deep speech prior. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(12), 2197-2212.

[5] Gelderblom, F. B., Tronstad, T. V., & Viggen, E. M. (2018). Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(3), 583-594.

[6] Yuliani, A. R., Amri, M. F., Suryawati, E., Ramdan, A., & Pardede, H. F. (2021). Speech Enhancement Using Deep Learning Methods: A Review. *Jurnal Elektronika dan Telekomunikasi*, *21*(1), 19-26.ff

[7] Han, W., Zhang, X., Min, G., Zhou, X., & Zhang, W. (2016, June). Perceptual weighting deep neural networks for single-channel speech enhancement. In *2016 12th World Congress on Intelligent Control and Automation (WCICA)* (pp. 446-450). IEEE.

[8] Jensen, J., & Taal, C. H. (2016). An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(11), 2009-2022.

[9] Gao, T., Du, J., Dai, L. R., & Lee, C. H. (2018, April). Densely connected progressive learning for lstm-based speech enhancement. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5054-5058). IEEE.

[10] Sun, L., Du, J., Dai, L. R., & Lee, C. H. (2017, March). Multiple-target deep learning for LSTM-RNN based speech enhancement. In *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)* (pp. 136-140). IEEE.

[11] Chern, A., Lai, Y. H., Chang, Y. P., Tsao, Y., Chang, R. Y., & Chang, H. W. (2017). A smartphone-based multi-functional hearing assistive system to facilitate speech recognition in the classroom. *IEEE Access*, *5*, 10339-10351.

[12] Zhao, Z., Liu, H., & Fingscheidt, T. (2018). Convolutional neural networks to enhance coded speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(4), 663-678.

[13] Islam, M., & Sobhan, M. (2014). BPN based likelihood ratio score fusion for audio-visual speaker identification in response to noise. *International Scholarly Research Notices*, *2014*.

[14] Strake, M., Defraene, B., Fluyt, K., Tirry, W., & Fingscheidt, T. (2020). Speech enhancement by LSTM-based noise suppression followed by CNN-based speech restoration. *EURASIP Journal on Advances in Signal Processing*, *2020*(1), 1-26.

[15] Cheng, J., Liang, R., & Zhao, L. (2020). DNN-based speech enhancement with self-attention on feature dimension. *Multimedia Tools and Applications*, *79*(43), 32449-32470.

LIST OF PUBLICATIONS

INTERNATIONAL CONFERENCE

- Presented a Paper titled “Role of Deep Neural Network in Speech Enhancement: A Review” in SLAAI International Conference on Artificial Intelligence (ICAI – 2018) in University of Moratuwa, Columbo, Sri Lanka. DOI - 10.1007/978-981-13-9129-3_8 **[Published in Communications in Computer and Information Science and Indexed in Scopus]**

- ❖ Presented a Paper titled “Enhancing the Quality and Intelligibility of Speech using Deep Neural Networks”, in International Conference on Automation, Signal Processing, Instrumentation and Control (ICASIC – 2020) at Vellore Institute of Technology, India.

INTERNATIONAL JOURNAL

1. Hepsiba, D., & Justin, J. (2021). Enhancement of single channel speech quality and intelligibility in multiple noise conditions using wiener filter and deep CNN. Soft Computing, 1-11. doi : 10.1007/s00500-021-06291-2 – **Published and Indexed in Scopus [Impact Factor: 3.642]**

2. Hepsiba, D., & Justin, J. (2022). Computational Intelligence for Speech Enhancement using Deep Neural Network. Computer Assisted Methods in Engineering and Science. – [**Published]**

3. Hepsiba. D., & Justin, J. Improved Speech Quality and Discernibility under multiple noise conditions with LSTM and BPN – Communicated to Soft Computing.

4. Hepsiba, D., & Justin, J. Enhancement of Alaryngeal Speech in Multiple Noise Conditions using Deep Learning Models – Biomedical Signal Processing and Control – Under Review.